DOCUMENT RESUME

ED 260 082 TM 850 163

AUTHOR Barter, Alice K.; And Others
TITLE A Comparison of Two Instruments for Evaluating Composition.
PUB DATE [80]
NOTE 90p.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS College Entrance Examinations; Comparative Testing; *Essay Tests; Evaluation Criteria; Higher Education; *Interrater Reliability; Mastery Tests; *Rating Scales; Scoring; *Scoring Formulas; Test Manuals; Test Reliability; *Writing Evaluation

ABSTRACT

A follow-up study of two instruments for evaluating college writing was conducted. The experimental scale (E Scale) was developed in 1976 and revised for this study. The control scale (C Scale) was described in the literature in 1977. Ten English majors graded ten essays from diagnostic entrance exams. Both the E Scale and the C Scale were used, with a six-week interval between evaluations. Analysis of variables such as mean scoring time, number of Pass and Fail papers, relationship of parts of the instruments to each other, frequency of disagreement between ratings, and evaluator's comments after each rating yielded significant results. The graders agreed with their initial assessment 78 percent of the time. The failure rate was higher on the E Scale; most grade changes were from Pass on the C Scale to Fail on the E Scale. With the E Scale, grading was about three minutes faster, but this difference was not significant. Evaluators preferred to use an instrument in grading rather than to use no instrument. Overall, the results of this follow-up were consistent with the original comparative study. Both scales are appended, as well as graders' comments and a summary of the original study. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED260082

A COMPARISON OF TWO INSTRUMENTS FOR EVALUATING COMPOSITION

Authors:

Alice K. Barter, Chicago State University
James A. Doppke
Arthur R. Parham

CS 850 163

BEST COPY

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. Barter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

TITLE: A COMPARISON OF TWO INSTRUMENTS FOR EVALUATING COMPOSITION

A pilot study and a follow-up study comparing two instruments for evaluating composition were conducted at Chicago State University. In the second study, ten English majors, nine graduate and one undergraduate, graded ten essays selected at random from a set of diagnostic entrance exams, each using two different instruments, the E-Scale and the C-Scale, at six week intervals between the first grading session and the second. Analysis of variables such as mean time, the number of Pass and Fail papers, the relationship of parts of the instruments to each other, the frequency of disagreement between a greader's first and second judgment, and others, in addition to the graders' responses to questionnaires after each of two trials, yielded some significant results. For the grader and/or student population in the second study the following obtained:

1. The graders agreed with their initial judgment 78% of the time.

2. The rate of failure was higher on the E-Scale than on the C-Scale; most changes of grade were from Pass on the C-Scale to Fail on the E-Scale.

3. The scales differed in the extent and significance of the parts to whole relationships within an instrument, the relationship of the parts to each other of the different instruments, the relationship of the parts to the overall Pass or Fail judgment, and the percentage of failure.

4. Though, on the average, grading by the E-Scale was 3.18 minutes faster than by the C-Scale, the difference in time did not prove to be significant.

5. Graders preferred to use an instrument to guide them in the grading of composition rather than to use no instrument.

6. On the whole, results of the second study were consistent with those of the pilot study.

TABLE OF CONTENTS

Page

LIST OF TABLES

SUMMARY OF THE MONOGRAPH

REPORT OF THE STUDY:

Introduction
Description of Study
Summary and Interpretations
Conclusion

APPENDICES:

I. Tables 1-15
II. Sample of E-Scale
III. Sample of C-Scale
IV. Grading Procedures for C-Scale
V. Form: Grader's Report, Trial 1, E-Scale Only, Part I
VI. Form: Grader's Comments, Trial, E-Scale and C-Scale
VII. Form: Grader's Comments, Trial, E-Scale and C-Scale
VIII. Report on Comparison of Two Instruments, Pilot Study, December, 1976
IX. Notes
X. References

LIST OF TABLES

Page

TABLE 1: Grader's Responses to Questionnaires, Trial 1 and Trial 2 (Part 1 Only)

TABLE 2: Graders' Comments, Trial 1 and Trial 2

TABLE 3: Means and Standard Deviations for Entire Sample

TABLE 4: Means and Standard Deviation Sorted by Pass and Fail

TABLE 5: Total Number of Pass or Fail Papers, Trial 1 and Trial 2 Combined

TABLE 6: Agreement Between Each Grader's Pass or Fail Judgment with E-Scale and C-Scale

TABLE 7: Comparison of Pass/Fail Judgments on Both Scales

TABLE 8: Direction of Changes of Grade From Trial 1 to Trial 2

TABLE 9: Mean Grading Time by Grader and Trial

TABLE 10: Correlations Between Variables in Experimental and Control Scales

TABLE 11: Matrix of Variables in Experiment Instrument

TABLE 12: Matrix of Variables in Control Instrument

TABLE 13: Difference from Mean of Part-Scores of Pass Papers v.s. Fail Papers on Ten Variables

TABLE 14: Comparison of Part I and Final Pass/Fail Judgments

TABLE 15: Comparison of Pass/Fail Judgments of C-Scale and E-Scale, Part I

A COMPARISON OF TWO INSTRUMENTS FOR EVALUATING COMPOSITION

INTRODUCTION

Purposes of the Study

In a 1975 National Council of Teachers of English publication, (1) the question "Is English an unmeasurable subject?" is addressed. The answer begins with the following:

> Being able to get valid and accurate instruments to measure student achievement in English depends upon one's ability to specify what is to be measured and what will be accepted as evidence that the thing to be measured has been achieved. Specifying what purpose the test is to serve also guards against inappropriate use of the results of the measurement. Measuring people's uses of their own language is not like measuring height or weight; there is necessarily some inexactness, some room for personal judgment. Indeed, much of the measurement is guess work. We estimate how a student thinks or feels from what the student says or writes, and we cannot be sure our estimate is correct.

In another NCTE publication, (2) Braddock notes:

> Unless one wishes to restrict one's conception of "writing" to the kind of essays employed in classrooms largely to test a student's understanding of a literary text or proficiency in avoiding certain "errors" covered by an English handbook, one sees writing as a staggeringly complicated and varied process.
>
> When "writing" is viewed as a varied and complicated concept, one readily sees that "writing ability" is a pair of words which has

> little meaning when taken in the abstract. Are there definable characteristics, understandings, and skills which are common to all modes of writing, for all kinds of prospective readers, for all possible purposes of writers? How can one consider "writing ability" in general when no one knows what it is? It may not even exist.

Indeed, the difficulty of finding valid instruments for measuring competence in English, especially in writing, is a concern not only of leaders of professional organizations, (3) researchers, and experts inthe field of test writing, (4) but also of the classroom instructor who must rely, sometimes to an extraordinary degree, on the judgments derived from their use.

The experimental instrument tested in this study was initially developed by the first author because of dissatisfaction with the instrument used for grading diagnostic and proficiency examinations at the urban university where she was teaching. It was first used in a pilot study there in 1976. Based on what was learned from this initial grading experiment, both the experimental scale and experimental methods were redesigned for the study being reported here.

The purposes of this study are as follows:

1) To measure the consistency of a grader's Pass or Fail judgment of the same essay when a) two different grading instruments are used and b) there is a time lapse of six weeks between the first and second grading;

2) To determine a grader's preference for either scale or no measuring instrument, according to stated criteria;
3) To gather data about the parts of the experimental instrument when it is used as an alternative to the control instrument.

The instrument currently in use has been previously reported in the literature.(5) In the discussion that follows it will be referred to as the C-Scale and the experimental instrument will be referred to as the E-Scale.

Discussions in recent literature (6) of evaluation theory attempt to analyze and classify theories about evaluating writing. Neither scale used in this study resembles the instruments developed from these theories: The ETS General Impression Procedure, the Diederich Analytic Scale, the Hirsch Analysis of Relative Reliability, the Lloyd-Jones Primary Trait Scoring, and M.A. Halliday's three function linguistic theory. However, the assumptions for E-Scale, Part I, appear to resemble the Diederich theory and those of the C-Scale, Dimension 1, the Hirsch theory. Also, the textual function portion of Halliday's theory seems to fit both Parts I and II of the E-Scale and Dimensions 1,2 and 4 of the C-Scale. Samples of the E-Scale and the C-Scale are shown in the appendices.

No attempt was made to examine the predictive validity of either scale. The final computer processing, using

transformations of scores based on the Rasch mathematical model, usually used with the C-Scale, was not performed. The current study focuses on the individual judgments of graders, using both the C-Scale and the E-Scale to evaluate each essay, rather than judgments made about the same essay by different graders. The degree of agreement between a grader's two judgments about the same essay is an important subject in the present study and would be obscured by transformations which attempt to make judgments of the graders uniform. An appropriate data transformation technique for the E-Scale which reduces inter-rater variability will be the subject for subsequent study. The practical limitations of the professional environment made elaborate training programs of graders impractical.

DESCRIPTION OF THE STUDY

In order to clarify what follows, definitions of key terms are given here:

1. Reliability:

   The .93 reliability of the C-Scale as reported in the literature is an inter-rater reliability. Since no such tests have been conducted to date for the E-Scale, the observations of agreement between judgments in Trial 1 and Trial 2 are evidence only of each grader's consistency. The usual methods of computing test reliability such as the split-half of Kuder-Richardson Formula 20 do not appear to be appropriate here. These methods assume a test made up of a fairly large number of similar questions. Retest reliability or rescoring reliability needs to be different here. If we are concerned with rescoring reliability, it could be done by having each rater rescore the same test, but retest reliability would almost be impossible to measure.

2. Trial

Trial is a scoring session. There was a time limitation of one week as graders graded the essays separately and returned them to class the following week. There was a six week interval between Trials 1 and 2.

3. Essay

These were 300 word essays written by entering freshmen on one of two questions distributed to them at the time of the exam. The sample of 100 essays was drawn from papers written on August 19, September 18 and September 13, 1978. The topics, in question form, were the following:

August 19:

Should parents of students convicted of vandalism be held responsible for their children's actions?

Should children have a say in how their abusive parents are punished?

September 18:

Should couples live together before getting married?

Should professionals, such as doctors, lawyers, and teachers, advertise their services?

September 19:

Should eighteen year olds be allowed to buy liquor?

Should interracial marriages be encouraged?

4. Dimensions

This is a term used for the parts of the "observation framework" in the Hake model. It is linked to and intends to represent linguistic categories:

"We believe that the observation framework now contains an exhaustive list of possible flaws for the types of essays we have graded. Our dimension set up is designed to reflect the generating process for the writer and provide cues to the reader. Thus in addition to presenting graders with criteria for judging the essay, it provides teachers a guide for teaching."

5. Agreement

Agreement as used here is the consistency of a grader's judgments of Pass or Fail in Trial 1 and Trial 2; it is the agreement of a grader with himself in grading the same essay six weeks after the first judgments and using a different measuring device.

6. Judgment

This is the grader's numerical (partial) judgments as reflected in the part-scores of both scales or the overall, qualitative judgment as represented by "Pass" or "Fail" on either scale.

7. Variable

It is any factor capable of being counted and affecting the results of this study. Variables of interest in this study were the number of students (100), the number of essays (100); the number of graders (10); the instruments (the C-Scale and the E-Scale); grading time (in minutes); and the number of Pass or Fail judgments.

Questions to be answered by this study are as follows:

1. Do the two instruments lead to judgments which are in substantial agreement?
2. Is the experimental instrument less time-consuming for the grader to use than the control instrument?
3. Does the judgment made following the use of the first part of the experimental instrument agree with that made following the use of the second part?
4. What are the graders' reactions as to the clarity, adequacy and efficiency of the two instruments?
5. Are the findings consistent with those of the pilot study?

Grading Procedures:

Ten graders were used. Each grader was randomly assigned ten papers to evaluate. Each grader evaluated the same paper with both instruments. In an attempt to avoid biasing the second evaluation, the evaluations with the two instruments were scheduled six weeks apart. Each grader used only one method of evaluation at a time. The order of the two evaluations was balanced: Five(one-half) of the graders were randomly chosen to use the experimental instrument during the first grading session with the remainder of the graders using the other instrument during that grading. For the second grading six weeks later, each grader used the instrument he or she had not used during the first grading session.

The ten graders were all English majors enrolled in an upper-level course in literary criticism; nine were graduate students and one was an undergraduate honor student. Though seven were high school teachers of English, none had had previous experience in grading with either scale in the study; it is for this reason that they were chosen. Each grader was paid a dollar per reading, that is, twenty dollars.

During the first half-hour of two class periods, six weeks apart, the authors conducted orientation sessions a half hour in length during which the graders were given a packet of ten composition exam papers, graders' instruction sheets for the instrument to be used, another sheet on which to record data and responses to the instrument. Instructions to the graders were as follows:

1. To grade the papers out of class, being careful to follow the directions exactly as given in the grading instrument.
2. Immediately after grading each paper, to record on the form provided the time in minutes that it took to grade the essay.
3. To answer a questionnaire on the adequacy, clarity and efficiency of the grading instrument and to offer suggestions for revising it.
4. To return the graded papers, the grading report form, the grading instrument and the questionnaire to their instructor the following week. (Copies of all these forms are in the appendices.)

At no time was the purpose of the study discussed with the graders. There was no practice session nor group grading session. The time in the orientation sessions was spent in teaching the graders how to fill out the forms as well as how to use the grading instruments.

The Instruments

The total number of possible errors measured by each grading instrument was as follows: C-Scale, Dimension 1 = 10; Dimensions 2+3+4 = unbounded; E-Scale, Part I = 18; Part II = unbounded. While cut scores have now been established for the C-Scale, none was used for either scale in this experiment. The version of the C-Scale that was used contained the following parts and titles:

Dimension 1 - The Exposition as a Whole: Logic, Organization, and Development

This dimension contains eight items in statement form. The total number of possible flaws or errors in this dimension is ten.

Dimension 2 - Meaning and Style

This dimension contains ten items in statement form. There are an unspecified number of possible errors in paragraphing, sentence structure (including fragments and run-ons), coherence, and diction.

Dimension 3 - Usage

This dimension consists of sixteen items in phrase form. There are an unlimited number of possible errors in verb, pronoun, noun, adjective/adverb usage, misspelled or misused words.

Dimension 4 - Punctuation

A list naming thirteen items makes up this dimension. There are an unlimited number of possible errors in the use of punctuation marks, underlining and capitalization.

A grader indicates each error by blackening the appropriate space on a test sheet in columns running alongside the student's lined essay paper.

The E-Scale was developed, pilot tested and modified by the first author. The intent of Part I of this scale is to go beyond the scope of Dimension 1 of the C-Scale, to define and assess the quality of a writer's thinking. It consists of the following:

A. Organization: Three questions about the controlling idea, the order of paragraphs, and pattern of introduction, body and conclusion.

B. Logical Development: Six questions about logical fallacies, such as unsupported generalizations, false analogies, irrelevancies, etc.

C. Complexity: Three questions about shallowness, limited development, lack of detail, etc.

There are twelve items in all intended to define A, B, and C. Failure to meet the criterion implicit in each question constitutes an error. The total number of possible errors in each is equal to six, the penalty value for each item in A and C being 2 and for B being 1. The maximum number of possible errors in Part I is 18. Another feature of Part I that differs from directions for Dimension I on the C-Scale is that the grader is asked to make a Pass or Fail judgment twice, once at the conclusion of grading Part I and then at the end of Part II. The main purpose of this step was to provide data about Part I as a predictor of the final Pass or Fail judgments. In the training sessions the graders were cautioned not to change this judgment after they had graded Part II, regardless of differences, as the information about these differences was of value to the study.

Part II of the E-Scale consists of unstructured grading of errors in usage, punctuation, sentence structure, paragraphing, etc. The grader is asked only a) to circle each error at its location; b) to use the symbol [#] for errors in paragraphing; and c) to add the total number of errors in Part II to those for Part I. He is to record each separately in finding the total error count for the essay.

He then is asked to make a final Pass or Fail judgment.

Sample:

In order to find students who had had no prior experience with the C-Scale, essays written as diagnostic exams by 100 randomly selected incoming freshmen in the summer of 1978 were used. Before they were given to the graders in this study, two copies of each of these essays were xeroxed and numbered at the 100 and 200 levels respectively in order to differentiate between those used in Trial 1 and those used in Trial 2.

Materials:

The materials used in this study were a) the C-Scale and the E-Scale; b) instruction sheets; c) forms for recording time, error counts and judgments; and d) questionnaires for graders. Samples of these are included in the appendices.

Findings:

Graders' responses to the questionnaires used at the end of Trial 1 and 2 yielded much information. The questionnaire used for Trial 1 stated, "An important part of this study is the grader's reaction to the instruments he is using to evaluate the papers. Please write as fully as you can your appraisal of the following aspects of the instrument you used in Trial 1, after answering each question with a 'yes' or 'no.'" The questions were directed toward the adequacy, clarity, and efficiency of the Trial 1 instrument. The Trial 2

questionnaire stated, "This form is to be completed at the conclusion of Trial 2. In the first portion, you will be asked to evaluate the instrument you used in Trial 2 in the same way you did the instrument you used for Trial 1. In the second half, you will be asked to make some additional judgments. Be sure you complete this portion as well." Part 1 of the Trial 2 questionnaire asked the same questions as the Trial 1 questionnaire except for the inclusion of one additional question, "Do you think it was easier to grade with this instrument than the instrument you used in Trial 1?" (See Table 1) Part II asked specific questions about suggestions for revisions, the grader's choice of instrument, and comparisons of the Trial 1 scale with the Trial 2 scale.

The data did not support the assumption that graders would prefer a scale that asks specific questions about the content of an essay, perhaps because the question being asked was not precise enough to elicit a usable answer to that question.

The responses in Table 1 show that after both trials more graders found the E-Scale easier to use than the C-Scale. The comments reported in Table 2 give mixed reactions to these two scales and suggest that both are wanting in one element or another that the grader is looking for.

The answers to Part II of the questionnaire were generally in paragraph form and were a part of the follow-up comments, some of which went beyond the specific questions asked.

The responses to question 1 show that both scales were found adequate and judgments were very close. However, Question 2, "Which of the following choices would you prefer for evaluating a set of compositions?" was easily tallied:

a. From those using the C-Scale for Trial 2:

Preference for E-Scale	2
Preference for C-Scale	2
No instrument	0
An instrument other than E-or C-Scales	0
No Response	1

b. From those using the E-Scale for Trial 2:

Preference for E-Scale	2
Preference for C-Scale	2
No instrument	1
An instrument other than E-or C-Scales	0
No Response	0

The responses to this question seemed to show that the order in which the instruments were used had little effect on a grader's preference. Responses to Question 3 show that there were fewer problems with the clarity of the directions for using the E-Scale than the C-Scale. The comments in Table 2, however, seem to indicate that, other than their own perceptions of what a grading instrument should do, the graders in this study had no bias in favor of either scale prior to its use. In general, the virtues that they saw in the E-Scale were its flexibility and ease of use. Its shortcomings, as perceived by some graders, were faults of omission such as lack of detail

in Part II. The C-Scale was cited for its precision in locating errors, but most graders found it more time-consuming and difficult to use.

The means and standard deviations for fourteen variables are given in Table 3 for the entire sample. Table 4 gives the means and standard deviations for twenty-six variables when the sample was sorted into those who passed and those who failed. Note that in Table 4 those who passed with the C-Scale had a higher mean error count in Dimension 4 and in the total error count for Dimension 2,3, and 4 combined than did those who failed. Table 5 shows that more students passed with the C-Scale than the E-Scale.

The data and tests of significance (see Tables 6,7) indicate that there was no significant change in the grader's Pass or Fail judgments from Trial 1 to Trial 2. There was 78% agreement between a grader's Pass or Fail judgments in Trial 1 and Trial 2. (See Table 7) A chi-square test showed this agreement between judgments on the two trials was significant at the 0.0001 level.

The total number of changes of grade from Trial 1 to Trial 2 was twenty-two. The percent of change overall was 0.11. The mean number of changes of grade per grader was 2.20, about equivalent to the results in the pilot study. Table 8 shows the directions of the changes of grade. The most frequent change of grade was from Pass with the C-Scale, Trial 1,

to Fail with the E-Scale, Trial 2, that is, when the E-Scale was the second instrument used.

An analysis of the differences in time required to grade with each scale revealed a significant difference in the time used. (Table 9) Moreover, the grading time was 3.18 minutes less when the E-Scale was used. The mean time for the E-Scale (12.10) is higher in Trial 2 than in Trial 1 (10.80), while on the C-Scale it is less in Trial 2 (13.40) than in Trial 1 (15.72). Three of the ten graders used more time in Trial 2 than in Trial 1. This finding suggests that these slow graders were themselves a factor in the time differences between Trials 1 and 2. The relationship between the Pass or Fail judgment and grading time was not signficant for either scale.

Table 10 shows that for some variables there was a significant positive relationship between the parts of the two instruments but not others. The table also shows that, except for E-Scale, Part I - Pass, and C-Scale, Dimension 1, there was no significant positive relationship between the part-scores from a given instrument and the final Pass or Fail judgment.

Table 11 shows those parts of the E-Scale that have significant positive relationships to each other and Table 12 shows the significant positive relationships between the parts of the C-Scale. Table 13 gives the difference from the mean of the part-scores, comparing Pass papers to Fail papers. Table 14 and 15 show that the grader's judgment following use of Part I

of the E-Scale was a highly significant predictor of the final judgments made with either scale. The positive relationship between the mean scores of each of the thirteen variables and the final Pass or Fail judgment was significant only for C-Scale, Dimension 1, and for the following ratings of the E-Scale: Part I, A,B,C; Total Error Count, Part I; and Total Error Count, Part II. With C-Scale, Dimension 2, there was little difference between the mean error counts of those who passed and those who failed. The mean error count with the C-Scale when the scores for Dimensions 2,3, and 4 were combined and also for Dimension 1 alone were slightly higher for those who passed than for those who failed.

Summary and Interpretations

The data from this study and their implications may be summarized as follows:

1. In spite of the time interval between the two trials, there was a high percentage of consistency between a grader's judgments on the two trials. The implications here may be a) the scales are measuring similar elements of writing; b) a grader's values or standards obtained regardless of the instrument used; or c) some factor not yet tested is operating.

2. The score at the end of Part I of the E-Scale is a good indicator of whether an essay will be passed or failed, regardless of which of the two scales, the E-Scale or the C-Scale, is used.

3. In terms of those who passed and those who failed, the most signficant differences occurred with Dimension 1 of the C-Scale and with these measures of the E-Scale: Part I, Total Error Count; Part II, Total Error Count, and Total Error Count, Part I plus Part II. These results suggest that these are the most discriminating items of each scale.

4. Most cases of disagreement between the two judgments occurred between Pass with the C-Scale and Fail on the E-Scale. One implication of this finding might be that the second scale is adding a new "dimension" to the grading process that is not provided on the C-Scale. A second interpretation might be that since the experimental instrument is less structured and less detailed in its notation of errors, the grader is free to apply his own concepts of what a good essay should be. On the other hand, it would seem that since it has fewer specific items the E-Scale would have the greater percentage of students passing rather than the C-Scale. The data did not support the latter premise in either study.

5. There was a significant positive relationship between C-Dimension 1 and all parts of E-Part I, as indicated by the scores, but not between C-Dimension 2 and these parts. An explanation might be that both C-Dimension 1 and E-Part I measure competence in organization, while C-Dimension 2 is a grab-bag which includes paragraphing, sentence structure, diction, and other errors that on the E-Scale are marked in Part II.

6. There was a signficant positive relationship between C-Dimension 3 and E-Error Count, Part II, and C-Dimension 3 and E-Total Error Count. Significant relationships also were shown between C-Dimension 4 and both E-Total Error Count, Part II, and E-Total Error Count. Since D-Dimensions 3 and 4 and E-Part II measure errors in mechanics and usage, the relationship between the two is predictable. The relationship between each of these dimensions and the total error count on the E-Scale is less easily explained since Part I of the E-Scale has no items referring to usage or mechanics.

7. There was a mean difference of 3.18 minutes in favor of the E-Scale in the time that it took to grade an essay with each instrument. Analysis of the order in which the instruments were used showed that graders using the E-Scale for Trial 2 took a longer period of time (12.10 vs. 10.80) than the Trial 1, E-Scale, graders, and those using the C-Scale for Trial 2 took a shorter period of time (13.70 vs. 15.72) than those using the C-Scale for Trial 1. Perhaps it is the grader's speed as much as or more than the order of the use of the instruments that is being indicated by the results.

8. Graders prefer an instrument to guide them in the evaluation of compositions. This finding is consistent with earlier studies and is not surprising since most teachers are aware of the disagreements and confusion within the profession about the assessment of writing.

9. The graders' responses to the two instruments showed that there was no clear preference of one scale over the other and that it made no difference to the graders which scale was used first or second.

   The premise that the order of use would prejudice the reader in favor of the first scale was not supported by these graders' responses. In the pilot study when only the C-Scale was used in Trial 1 and only the E-Scale in Trial 2, there was a preference stated for the C-Scale. Also all of the graders in the pilot study had had considerable experience in using the C-Scale.

10. The rate of failure was higher when graders used the E-Scale than when they used the C-Scale, suggesting either that the E-Scale is more severe or that it is measuring something the C-Scale is not. Nevertheless, a weakness in the design of this study is that there is no clear evidence that the differences in the results from the two scales are related to intent of Part I of the E-Scale, which is to measure the quality of content of an essay. If the study were repeated, data related to this factor should be obtained. One obvious, though superficial, means of doing so would be the graders' questionnaires.

Conclusion

The substantial agreement between the results from the two grading instruments compared in this study indicates that the experimental instrument should be investigated further as an alternative to the control instrument in cases where detailed diagnostic information is not required. The development of an appropriate data transformation technique to reduce inter-rater variability should be a major goal of subsequent investigation. Another idea for further research might be to

combine Part I of the E-Scale with Dimensions 2,3, and 4 of the C-Scale to create a new scale. The results from this hybrid instrument should be compared with the results reported here. Still a third area for research might be to pursue the question of the value to the teacher of a grading instrument for essays by replicating this study but using no instrument for Trial 1.

Finally, to return to the question of whether writing is a measurable subject, we must answer that it is certainly a difficult but not an impossible one to measure. As a review of the literature on the evaluation of writing continues to disclose, evaluation is a stormy sea of uncertain depths and shallows.

I. Tables 1-15

TABLE 1
GRADERS' RESPONSES TO QUESTIONNAIRES
TRIAL 1 AND TRIAL 2 (PART I ONLY)

Question	Trial 1 C-Scale	Trial 1 E-Scale	Trial 2 C-Scale	Trial 2 E-Scale
1. Adequacy for measuring competance in composition:				
a. Does it meet with your standards for a logical, coherent, reliable tool for evaluating composition?				
yes——	5	5	4	5
no——	0	0	1	0
b. Does it provide for the qualities you expect in good writing?				
yes——	5	4	4	5
no——	0	1	1	0
c. Is there anything omitted that should be included?				
yes——	1	2	2	2
no——	3	3	2	3
no response	1	0	1	0
d. Is it properly weighted in its parts?				
yes——	2	4	4	4
no——	1	0	1	1
no response—	1	1	0	0
not sure——	1	0	0	0

BEST COPY AVAILABLE

BEST COPY AVAILABLE

TABLE 1 (Con't)

2. Clarity of the directions for using the instrument:				
a. Is there any portion of this instrument that is unclear or ambiguous?				
yes----	5	3	2	1
no-----	0	2	3	4
3. Efficiency and ease in evaluating a set of papers:				
a. Was the instrument easy to use?				
yes----	3	4	2	5
no-----	2	1	3	0
b. Do you think it was easier to grade a paper with this instrument than with no instrument?				
yes----	3	4	3	4
no-----	2	1	2	1
c. Do you think using the instrument decreased the time you would normally take to grade a paper?				
yes----	2	4	1	2
no-----	2	0	4	3
no response---	1	1	0	0
d. Do you think using the instrument increased the time you would normally take to grade a paper?				
yes----	2	0	4	3
no-----	2	4	1	2
no response---	1	1	0	0
e. (Trial 2 only) Do you think it was easier to grade with this instrument than with the instrument you used in Trial 1?				

yes———	1	4
no———	4	1

TABLE 2
GRADERS' COMMENTS, TRIAL 1 AND TRIAL 2

Grader No.	Trial 1 Scale	Comments	Trial 2 Scale	Comments
3	E-Scale	The term question-begging is unclear. Following the form made papers quicker to grade	C-Scale	Ithink I was faster, but I was clearer and more definite.
4	E-Scale	Some emphasis should be made on clarity of thought. Penmanship was very bad in most cases. Most topics were too difficult for freshman students.	C-Scale	It is not good for an essay type test. Compositions are too short for test accuracy.
6	E-Scale	It provides for qualities expected in good writing (in terms of content), but criteria for evaluating mechanics should be included. I would like more grading alternatives than Pass or Fail. It is not clear how many errors or what kinds of errors merit failure. It should provide some way to classify the errors; combine it with a tool for evaluating mechanics. I prefer this test to no instrument because it did provide structure.	C-Scale	It provides for qualities of good writing, but not for shallowness of immaturity. This test seems to be more concise.

7	E-Scale	No Comment.	C-Scale	No Comment.

Grader No.	Trial 1 Scale	Comments	Trial 2 Scale	Comments
8	E-Scale	I think it was superb.	C-Scale	It should simplify the process of error-making. It is effective but too time consuming. It is too complex and time-consuming, but the Trial 1 test is too simplistic.
2	C-Scale	I like the idea and organization of the dimension method of grading; it is a step toward standardized grading of essays. Some improvements seem in order. Some ele- of grammar, such as the correct format of plurals, are not included. Some items need further explanation: It is not clear to me whether "No Body" (D1, 5) means only an introduction and conclusion are present, or whether no paragraph divisions indicate "No Body."	E-Scale	No Comment.
5	C-Scale	After reading each paper, I found it easier to correct each error as I found it. Depending on the grade level of the student, I would prefer to make the exact correction rather than to indicate just the dimension.	E-Scale	It was easier for me to use this instrument because the errors did not have to be shaded in as they were for the first trial. With this instrument, I wasn't quite sure how to deal with incomplete or ambiguous sentences.

Dimension 1 was the most difficult section to correct on the papers. There were many usage and punctuation errors. Passing and failing could depend upon the grader's point of emphasis. Many public schools students are involved in reading mastery and creative writing. Less emphasis is placed on usage, mechanics, etc.

I found that I was a little more severe in grading with this instrument. The first trial was more detailed, and I found myself having to refer to the different sheets quite often. Due to a lack of time, I think I would prefer this last instrument.

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Grader No.	Trial 1 Scale	Comments	Trial 2 Scale	Comments
9	C-Scale	This grading system was helpful in the discovery of grammatical errors. However, I found it a bit difficult to grade overall Pass or Fail.	E-Scale	I think the Trial 1 test was too detailed in its evaluation of the grammatical area of the test. The error codes were somewhat ambiguous. I believe that the Trial 2 test was very good. It allowed for more leeway on the part of the grader. I think it is necessary to have some sort of instrument for grading. I think I would prefer the Trial 2 test. It has very good but general outlines for grading. Although not all of the errors may be caught, there is less ambiguity in marking errors. The strengths of Parts I and II are the giving of guidelines for grading. The grader has some idea of what to look for while reading the essay. Its weaknesses lie in the time element for grading Part I and in the uncertainty of marking of errors. There are some errors not listed, and the grader has some trouble in marking.
1	C-Scale	Everyone who grades papers has something he or she looks for. Dimension 1 was unclear and so was Dimension 2.	E-Scale	If a grader is given a standard, then instruments would not be necessary.

		I feel that graders should be told what a passing paper should be and then let them grade.		
Grader No.	Trial 1 Scale	Comments	Trial 2 Scale	Comments
10	C-Scale	I found this to be a very difficult assignment, probably because I am a non-teaching major. For this reason, I found this instrument both a godsend for giving evaluation standards and a pain in the neck because I am unskilled in the kind of analysis this evaluation required. Since I am unsure how to weigh the various dimensions, I decided that I and II represented critical components of good writing and based my Pass or Fail judgments on them. I was uncertain whether to label some errors as Dimension II defective sentences or Dimension III usage errors. The directions for using the instrument were unclear here. Therefore, if the sentence was generally correct, I labeled the mistakes as usage errors; if the sentence was badly flawed, I marked it for a Dimension II error.	E-Scale	Please remember that I do not teach and have no experience grading papers. My only suggestion for revising Test 1 would be to assign weights to the various dimensions to classify Pass or Fail judgments. I remember arbitrarily weighting the dimension dealing with the structure as the heaviest. I believe if you can't organize a coherent paper with an introduction, body and conclusion, you can't write essays. Test 2 needs a category to deal with such problems as sentence fragments, run-on sentences, non-parallel structure, etc. Even though I found the first instrument difficult to use because it was so detailed, I now find that I prefer it for that very reason. If I had more grading experience, I might not need so much structure in an instrument. Test 1: Strengths: detailed checklist; a good blueprint for evaluating essays.

Weaknesses: difficult to use; lack of weighting of various dimensions.

Test 2:

Strengths: the section specifying what is meant by logical development.

Weaknesses: categories too general.

TABLE 3
MEANS AND STANDARD DEVIATIONS FOR ENTIRE SAMPLE

Variable	N	Mean	Standard Deviation
E-Time	100	11.5400	4.7127
C-Time	100	14.6300	9.1845
EC-Dimension 1 (a)	100	2.1400	2.0351
EC-Dimension 2	100	5.4000	3.7254
EC-Dimension 3	100	7.1400	6.8843
EC-Dimension 4	100	4.4100	4.7185
EC-PIA (b)	100	2.2400	1.9905
EC-PIB	100	1.6300	1.5234
EC-PIC	100	2.1700	1.9491
TEC-PI (c)	100	6.1300	4.5431
EC-PII (d)	99	16.4444	10.9995
TEC	100	22.4100	11.5006
TEC Dimensions 2, 3, 4	100	16.9500	11.1504

(a) EC = error count
(b) PI = Part I (E-Test)
(c) TEC = total error count
(d) PII = Part II (E-Test)

TABLE 4
MEANS AND STANDARD DEVIATION SORTED BY PASS AND FAIL

Variable	N	Mean	Standard Deviation
C-Time (Fail)	36	14.5000	11.4330
C-Time (Pass)	64	14.7031	7.7412
EC-Dimension 1 (Fail)	36	3.3889	2.2204
EC-Dimension 1 (Pass)	64	1.4375	1.5417
EC-Dimension 2 (Fail)	36	5.3333	4.4013
EC-Dimension 2 (Pass)	64	5.4375	3.3232
EC-Dimension 3 (Fail)	36	8.1667	7.7108
EC-Dimension 3 (Pass)	64	6.5625	6.3643
EC-Dimension 4 (Fail)	36	3.3055	3.9915
EC-Dimension 4 (Pass)	64	5.0312	5.0047
TEC-Dimensions 2,3,4 (Fail)	36	16.8055	11.9996
TEC-Dimensions 2,3,4 (Pass)	64	17.0312	10.7408
E-Time (Fail)	46	12.3696	5.3598
E-Time (Pass)	54	10.6670	3.9668
EC-PIA (Fail)	46	3.3043	1.8958
EC-PIA (Pass)	54	1.3333	1.5901
EC-PIB (Fail)	46	2.4565	1.6827
EC-PIB (Pass)	54	1.0185	0.9808
EC-PIC (Fail)	46	3.3261	1.9782
EC-PIC (Pass)	54	1.1852	1.2749
TEC-PI (Fail)	46	9.1304	4.1612
TEC-PI (Pass)	54	3.5471	3.0629
EC-PII (Fail)	46	20.3111	11.9465
EC-PII (Pass)	54	13.2222	9.0442
TEC (Fail)	46	29.0000	11.0212
TEC (Pass)	54	16.7963	8.6037

Table 5: Total Number of Pass or Fail Papers,
Trials 1 and 2 Combined

Number of Pass Papers	Number of Fail Papers	Number Passing C-Scale but Failing E-Scale	Number Passing E-Scale but Failing C-Scale
E-Scale=23 C-Scale=43	E-Scale=71 C-Scale=51	16	9
Total=66	Total=112		

BEST COPY AVAILABLE

TABLE 6
AGREEMENT BETWEEN EACH GRADER'S PASS OR FAIL JUDGMENT
With E_SCALE AND C-SCALE

Grader	Judgment	Agreement N	Percent	Disagreement N	Total Changes Percent of Grade	
1	Fail	4	0.80	1	0.20	2
	Pass	4		1		
2	Fail	1	0.50	6	0.50	5
	Pass	4		0		
3	Fail	1	0.80	2	0.20	2
	Pass	7		0		
4	Fail	2	0.80	0	0.20	2
	Pass	6		2		
5	Fail	0	0.60	4	0.40	4
	Pass	6		0		
6	Fail	1	0.80	1	0.20	2
	Pass	7		1		
7	Fail	7	0.80	1	0.20	2
	Pass	1		1		
8	Fail	4	1.00	0	0.00	0
	Pass	6		0		
9	Fail	4	0.90	0	0.10	1
	Pass	5		1		
10	Fail	6	0.80	2	0.20	2
	Pass	2		0		

BEST COPY

TABLE 7

COMPARISON OF PASS/FAIL JUDGMENTS ON BOTH SCALES'

PASS C-PASS
C-SCALE

FREQUENCY PERCENT ROW PCT COL PCT	0	1	TOTAL
0	30 30.00 65.22 83.33	16 16.00 34.78 25.00	46 46.00
1	6 6.00 11.11 16.67	48 48.00 88.89 75.00	54 54.00
TOTAL	36 36.00	64 64.00	100 100.00

1
PASS- E-SCALE: C-PASS,C-SCALE

BEST COPY AVAILABLE

TABLE 8
DIRECTION OF CHANGES OF GRADE
FROM TRIAL 1 TO TRIAL 2

Trial	Scale	Pass	Fail
1	C-Scale	12	2
	E-Scale	4	4
2	C-Scale	4	4
	E-Scale	2	12

TABLE 9
MEAN GRADING TIME BY GRADER AND TRIAL

Grader	Trial 1 Method	Time[a](N=10)	Trial 2 Method	Time[a](N=10)	Differences in Time Between Trial 1 and Trial 2
1	C-Scale	20.50	E-Scale	17.90	-2.60
2	C-Scale	9.80	E-Scale	8.20	-1.60
3	E-Scale	9.50	C-Scale	11.40	+1.90
4	E-Scale	14.80	C-Scale	14.80	0.00
5	C-Scale	19.10	E-Scale	11.20	-7.90
6	E-Scale	12.80	C-Scale	12.60	-0.20
7	E-Scale	8.30	C-Scale	5.60	-2.70
8	E-Scale	8.60	C-Scale	22.20	+13.60
9	C-Scale	23.40	E-Scale	10.10	-13.30
10	C-Scale	5.80	E-Scale	13.10	+7.30

Mean Time	Trial 1	Trial 2
C-Scale	15.72(N=50)	13.40(N=50)
E-Scale	10.80(N=50)	12.10(N=50)

[a] In minutes.

TABLE 10

CORRELATIONS BETWEEN VARIABLES IN EXPERIMENTAL AND CONTROL SCALES[1]

	C-Scale Variables					
E-Scale Variables	D1	D2	D3	D4	C-TIME	C-PASS
ECPIA	0.45043	0.03324	0.00563	-0.15792	0.00491	-0.43507
N=100	*0.0001	0.7427	0.9557	0.1166	0.9614	*0.0001
ECPIB	0.29478	-0.11248	0.01876	-0.08133	0.10046	-0.41947
N=100	*0.0029	0.2652	0.8530	0.4211	0.3200	*0.0001
ECPIC	0.18237	-0.11657	-0.06502	-0.15043	-0.12736	-0.46063
N=100	0.0694	0.2481	0.5204	0.1352	0.2067	*0.0001
ECPII	0.10308	0.37055	0.61101	0.40491	0.13354	-0.15344
N=99	0.2869	*0.0002	*0.0001	*0.0001	0.1876	0.1295
E-TIME	0.14318	0.02647	0.31872	-0.05335	0.10400	-0.28790
N=100	0.1412	0.7958	*0.0012	0.5981	0.3051	*0.0037
PI-PASS/FAIL	-0.35616	-0.02906	0.03381	0.05810	0.11403	0.54715
N=100	*0.0003	0.7741	0.7385	0.5658	0.2586	*0.0001
E-PASS	-0.37217	0.05738	-0.21840	-0.00915	0.16023	0.56180
N=100	*0.0001	0.5707	0.0290	0.9280	0.1113	*0.0001

[1] Correlations are given on the first line and the level of significance on the second line for each variable. Significant levels are preceded by an asterisk.

TABLE 11

MATRIX OF VARIABLES IN EXPERIMENTAL INSTRUMENT

	ECPIA	ECPIB	ECPIC	ECPII	E-TIME	PI	PASS
ECPIA	1.00000	0.49360	0.49447	-0.05355	0.01960	-0.61828	-0.49501
	0.0000	*0.0001	*0.0001	0.6000	0.8465	*0.0001	*0.0001
ECPIB	0.49360	1.00000	0.56279	-0.04681	0.20457	-0.58538	-0.47282
	*0.0001	0.0000	*0.0001	0.6454	0.0412	*0.0001	*0.0001
ECPIC	0.49447	0.56279	1.00000	-0.06321	0.00588	-0.56294	-0.55019
	*0.0001	*0.0001	0.0001	0.5342	0.9537	*0.0001	*0.0001
ECPII	-0.05355	-0.04681	-0.06321	1.00000	0.31223	-0.04990	-0.32265
	0.6000	0.6454	0.5342	0.0000	*0.0017	0.6238	*0.0011
E-TIME	0.01960	0.20457	0.00583	0.31223	1.00000	-0.07664	-0.18100
	0.8465	*0.0412	0.9537	*0.0017	0.0000	0.4485	0.0715
PI	-0.61828	-0.58538	-0.56294	-0.04990	-0.07664	1.00000	0.74721
	*0.0001	*0.0001	*0.0001	0.6238	0.4485	0.0000	*0.0001
PASS	-0.49501	-0.47282	-0.55019	-0.32265	-0.18100	0.74721	1.00000
	*0.0001	*0.0001	*0.0001	*0.0011	0.0715	*0.0001	0.0000

[1] All correlations have N=100 except those involving ECPII(Error Count Part II) which have N= 99. Correlations are given in the first line and the level of significance on the second line for each variable. Significant levels are preceded by an asterisk.

TABLE 12

MATRIX OF VARIABLES IN CONTROL INSTRUMENT[1]

	D1	D2	D3	D4	GTIME	GPASS
D1	1.00000	0.07780	0.11754	-0.03654	-0.11933	-0.46256
	0.0000	0.4416	0.2441	0.7181	0.2370	*0.0001
D2	0.07780	1.00000	0.30933	0.35661	0.36128	0.01349
	0.4416	0.0000	*0.0017	*0.0003	*0.0002	0.8940
D3	0.11754	0.30933	1.00000	0.19069	0.25595	-0.11241
	0.2441	*0.0017	0.0000	0.0574	*0.0102	0.2655
D4	-0.03654	0.35661	0.19069	1.00000	0.09444	0.17643
	0.7181	*0.0003	0.0574	0.0000	0.3500	0.0791
GTIME	-0.11933	0.36128	0.25595	0.09444	1.00000	0.01067
	0.2370	*0.0002	*0.0102	0.3500	0.0000	0.9161
GPASS	-0.46256	0.01349	-0.11241	0.17643	0.01067	1.00000
	*0.0001	0.8940	0.2655	0.0791	0.9161	0.0000

[1] All correlations have N = 100. Correlations are given on the first line and the level of significance on the second line for each variable. Significant levels are preceded by an asterisk.

Table 13: Difference from Mean of Part-Scores of Pass Papers v.s. Fail Papers on Ten Variables

Variable	Error Count Range-	Mean-Variable	Mean-Pass	Difference from Mean-V	Mean-Fail	Difference from Mean-V
1. C -D 1	0-9	2.1400	1.44	-.70	3.39	+1.25
2. C -D 2	0-15	5.4000	5.44	+.04	5.33	+.07
3. C -D 3	0-28	7.1400	6.50	-.64	8.17	+1.03
4. C -D 4	0-29	4.4100	5.03	+.62	5.30	+1.11
5. E -Pt IA	0-6	2.2400	1.33	-.91	3.30	+1.06
6. E -Pt IB	0-6	1.6800	1.02	-.66	2.46	+.78
7. E -Pt IC	0-6	2.1700	1.18	-.99	3.33	+1.06
8. E -TEC-Pt I	0-18	6.1300	3.59	-2.54	9.13	+3.03
9. E -EC-Pt II	0-54	16.4400	13.22	-3.22	20.31	+3.87
10. E -TEC	3-58	22.4100	16.80	-5.61	29.00	+6.59

TABLE 14
COMPARISON OF PART I AND FINAL PASS/FAIL JUDGMENTS

Final Pass/Fail PI (Pass/Fail Judgment after Part I)

FREQUENCY PERCENT ROW PCT COL PCT	0	1	TOTAL
0	35 35.00 76.09 94.59	11 11.00 23.91 17.46	46 46.00
1	2 2.00 3.70 5.67	52 52.00 96.30 82.54	54 54.00
TOTAL	37 37.00	63 63.00	100 100.00

The Pass/Fail judgment made after the first part of the E-Scale agreed with the final judgment in the E-Scale in 87% of the cases.

BEST COPY AVAILABLE

TABLE 15
COMPARISON OF PASS/FAIL JUDGMENTS of C-SCALE AND E-SCALE, PART I

PI C-PASS

FREQUENCY PERCENT ROW PCT COL PCT	0	1	TOTAL
0	26 26.00 70.27 72.22	11 11.00 29.73 17.19	37 37.00
1	10 10.00 15.87 27.78	53 53.00 84.13 82.81	63 63.00
TOTAL	36 36.00	64 64.00	100 100.00

The Pass/Fail judgment made with the C-Scale was not independent of that made with the first part of the E-Scale as shown by the highly significant chi-square statistic (.0001). The Pass/Fail judgment made with the first part of the E-Scale agreed with the C-Scale judgment in 79% of the cases. The changes of judgment were almost evenly divided, indicating that the C-Scale and the first part of the E-Scale are of equal difficulty.

BEST COPY AVAILABLE

II. Sample of E-Scale

EVALUATION OF COMPOSITION

ESCALE FORM A

Description

ESCALE Form A is an instrument for evaluating composition. It is divided into two parts. In Part I thegrader is asked to evaluate the quality of the thinking in the essay. Three attributes--organization, logical development, and complexity--are the factors being measured. The questions given beneath each attribute are to be scored separately but in relation to the attribute specified; they are provided to help define the attribute for the grader.

In Part II the grader is asked to do the following:

1. To circle the location of each error in usage, spelling, punctuation, capitalization, sentence structure, etc.

2. To use the circled symbol ⓟ in the cases of errors in paragraphing, (a) at the point where a new paragraph should begin or (b) at the point where a new paragraph should not have begun. Each paragraphing fault should be counted as a single error and included in the cumulative count.

The scoring of Form-A items in Part I ranges from "0" to "-6". A zero score indicates satisfactory performance in a specific attribute, while a minus score indicates that the essay is flawed in this respect. To illustrate:

Part I	Yes (Penalty)	No (No Penalty)
1. Organization	(Range: -2 to -6)	0
2. Logical Development	(Range: -1 to -6)	0
3. Complexity	(Range: -2 to -6)	0

The scoring for Part II will be the cumulative error count for that part only.

After adding the error count for Parts I and II, the grader is to mark the paper "Pass" or "Fail" in the space provided at the top of Page 2.

BEST COPY

TEST - FORM A

Record the following information for each paper in the space provided below:

Student _____ Grader _____

Instructor _____ Total Score - Part I _____

Class or Grade Level _____ Total Score - Part II _____

Date _____ Total Score - Parts I & II _____

Final Grade (Pass or Fail) _____

Directions for Scoring Essay:

1. Part I--Read the essay once without marking any flaws. Now judge the essay as a whole in relation to the following questions. Give your scoring of each factor in the space to the right of the factor named. If, on the whole, the answer is No, record a zero in the Score Column. When the answer, on the whole is Yes, record the minus grade value indicated in the Score Column:

	Yes (Penalty)	No (No Penalty)	Score Column
b. Organization:			
(1) Does the writing fail to sustain a controlling idea?	(-2)	0	___
(2) Is the order of the paragraphs illogical or incoherent?	(-2)	0	___
(3) Is the writing lacking some part of the organizational pattern of introduction, body, and conclusion?	(-2)	0	___
b. Logical Development:			
Aside from logical order, does the essay have faulty logic such as:			
(1) Unsupported generalizations	(-1)	0	___
(2) Inaccurate information	(-1)	0	___
(3) Unsubstantiated or illogical cause/effect relationships	(-1)	0	___
(4) Irrelevancies	(-1)	0	___
(5) False analogies	(-1)	0	___
(6) Question begging	(-1)	0	___

BEST COPY

c. Complexity:

(1) Is the writing too shallow too immature, or too limited in its development of the central idea?	(-2)	0	___
(2) Does it lack concrete and relevant detail?	(-2)	0	___
(3) Does the writer fail to associate the main idea with other relevant ideas, facts or points of view?	(-2)	0	___

****STOP: 1. Record the total score for Part I in the space provided at the top of Page 2.

2. Now continue with Part II.

2. Part II--Reread the paper. On the paper circle the places on the line where error of any kind occurs. To illustrate::

Typing is for the birds ( ) it certainly is

not my (fort) (Lass) year I (pass) typing after

taking it twice before.

Count each error of faulty paragraphing as one point, using the symbol ¶ at the point where a new paragraph should begin or where a new paragraph should not have begun.

****STOP: 1. Record the total error count for Part II in the space provided (Page 2).

2. Add your total score for Parts I and II and record it in the space provided (Page 2).

3. If, in your judgment, the essay is a satisfactory essay as a whole, write "Pass" in the space provided on Page 2. If it is unsatisfactory, write "Fail" in the space provided.

4. Attach these sheets to each graded essay.

III. Sample of C-Scale

C-SCALE

The four-dimensional framework which follows is just that--a framework for measuring rudimentary competence in composition. The framework is built on a basic expositiory essay model (the simplest) and assumes the use of standard American English; it does not imply that this is the only possible model. The purpose is to provide a means for relatively systematic and specific recording of what the grader perceives as flaws in an exposition.

DIMENSION I

The Exposition as a Whole: Logic, Organization, and Development

The expositiory essay is flawed because

1. it does not address the topic and therefore is not eligible for grading,
2. it is too short and therefore is not eligible for grading,
3. it has no introduction (3 flaws),
4. it has a faulty introduction (e.g. no stated or implied central idea, does not place the central idea in an overall context, does not imply how the central idea is to be developed--1 flaw).
5. it has no body (5 flaws),
6. it has a faulty body
    a. does not develop the central idea (2 flaws),
    b. is not well organized (2 flaws),
    c. does not include sufficient detail (1 flaw),
7. it needs a conclusion and does not have one (2 flaws),
8. it has a faulty conclusion (e.g. does not restate the central idea, has information irrelevant or contradictory to the introduction and/or body (1 flaw)

The total number of possible flaws in this dimension is 10.

DIMENSION II
Meaning and Style

The essay's meaning or style is flawed because.

1. The title is omitted or inappropriate,
2. A new paragraph should/should not begin,
3. A necessary paragraph is omitted from the essay,
4. A faulty* paragraph is included in the essay,
5. A necessary sentence is omitted from the paragraph,
6. A faulty* sentence is included in the paragraph,
7. A necessary element (word or word grouping) is omitted from the sentence,
8. A faulty* element is included in the sentence,
9. There is a sentence fragment,
10. There is a run on sentence or comma splice,

*Faulty can refer to incorrect, illogical, inconsistent, irrelevant, misplaced, redundant, repetitious, or unclear.

BEST COPY

DIMENSION III

Usage

The essay's usage flaws are in

1. Verb usage
   a. an improper subject/verb agreement
   b. a verb phrase omitting a verb
   c. an incorrect verb ending or verb form
   d. an inconsistent tense, mood, or voice

2. Pronoun usage
   a. no antecedent for a pronoun
   b. pronoun not agreeing with its antecedent
   c. pronoun in incorrect case form

3. Noun usage
   a. incorrect plural form
   b. no ploural form
   c. incorrect possessive form
   d. no possessive form

4. Adjective usage
   a. incorrect comparative or superlative form
   b. no comparative or superlative
   c. adjective instead of adverb or vice versa

5. Misspelled words

6. Misused words

DIMENSION IV

Punctuation

An error has been made by the omission or incorrect usage of the following:

1. Capital Letters
2. Period (unless the period creates a sentence fragment marked in dimension II)
3. Question mark
4. Exclamation point
5. Comma (unless the comma creates a comma splice marked in dimension II)
6. Colon
7. Semicolon (unless the semicolon creates a fragment marked in dimension II)
8. Quotation marks
9. Dash
10. Underlining
11. Hyphen
12. Parenthesis
13. Apostrophe

IV. Grading Procedures for C-Scale

Grading Procedure - C-Scale

Specific Procedures

1. Read through the Dimensional Grading Flaw Sheets and keep them nearby for ready reference.

2. Place your grader number in the appropriate space in the top right corner of the first page of the examination you are about to read.

3. Note the four columns down the right edge of the paper. Dimension one flaws are to be recorded in the column nearest the text, and next to the row of numbers that extend down the page. Dimension two flaws are recorded in the next column, dimension three flaws in the third column, and dimension four flaws in the column farthest to the right.

4. Read through the essay quickly with dimension one in mind. If you have dimension one flaws to record, mark each in the space next to the number that corresponds to the flaw you are recording, as indentified by number on the Flaw Sheets, page 1--e.g.: "has a faulty introduction"

	D_1	D_2	D_3	D_4
1				
2				
3				
4				
5				

Note that all dimension one flaws are entered together on page 1 and that items 1-8 exhaust the possibilities.

5. Reread the essay for possible errors in dimensions two, three, and four. Circle each error in the text, and blacken the space nearest to where the error occurs--e.g.:

Many sports, has become unprophetable in recent years, football is an example.

(10)

D_1	D_2	D_3	D_4
		-	-
	-		

6. In the case of dimension <u>two</u> errors, please <u>also</u> enter, next to the error circled, the number from page 2 of the Flaw Sheets which best identifies the error (in this case "comma splice" is error no. 10).

7. Indicate your final judgment of the essay (P, N, II or whatever) on the Grader Response line in the upper right corner of the essay's first page.

8. Total the count of the errors you have recorded in each dimension and enter these totals on the Grader Response line next to your judgment in the following order:

   (a) For the Diagnostic Examination: III 5-12-10-6 or I 2-4-5-7
   Note: Category number first, then scores.

   (b) For the <u>Composition I Examination</u>: P 1-3-4-4 or N 1-6-10-9 II
   Note: P/N judgment first; then scores; then, in the case of N, category recommendation.

   (c) For the <u>Qualifying Examination</u>: P 0-2-3-4 or N 1-4-11-4

9. <u>For all examinations</u>: If the students appears to be a non-native speaker of English or a bilingual speaker, indicate this as <u>BI</u>? next to the judment and forward the names of these individuals to the Composition Coordinator.

V. Form: Grader's Report, Trial 1, E-Scale Only, Part I

Form Used for E-Scale Only

Grader_____

GRADER'S REPORT FORM-PART I

1 Name of Student	2 Judgment (Pass or Fail)	3 Total Error Count-Part I

VI. Form: Grader's Comments, Trial 1, E-Scale and C-Scale

Date _____

Name of Grader: _____

GRADER'S COMMENTS--TRIAL 1

***Note: An important part of this study is the grader's reaction to the instruments he is using to evaluate the papers. Please write as fully as you can your appraisal of the following aspects of the instrument you used in Trial 1, after answering each question with a "Yes" or "No".

1. Adequacy for measuring competence in composition:

   a. Does it meet your studards for a ological, coherent, reliable tool for evaluating composition?

      Yes _____

      No _____

   b. Does it provide for the qualities you expect in good writing?

      Yes _____

      No _____

   c. Is there anything omitted that should be included?

      Yes _____

      No _____

   d. Is it properly weighted in its parts?

      Yes _____

      No. _____

2. Clarity of the directions for using the instrument:

   Is there any portion of this instrument that is unclear or ambiguous?

   Yes _____

   No _____

3. Efficiency and ease in evaluating a set of papers:

   a. Was the instrument easy to use?

      Yes _____

      No _____

b. Do you think it was easier to grade the paper with this instrument <u>than with no</u> instrument?

Yes _____

No _____

c. Do you think using the instrument <u>decreased</u> the time you would normally take to grade a paper?

Yes _____

No _____

d. Do you think using the instrument increased the time you would normally take to grade a paper?

Yes _____

No _____

****Please write your comments on the back of these pages. If you use additional paper, please attach the sheets to this form.

Thanks.

VII. Form: Grader's Comments, Trial 2, E-Scale and C-Scale

Form Used for P-Scale and C-Scale

Date_____

Grader_____

GRADER'S COMMENTS --TRIAL 2

*Note: This form is to be completed at the conclusion of Trial 2.

In the first half you will be asked to evaluate the instrument you used in Trial 2 in the same way as you did the instrument you used for Trial 1. In the second half you will be asked to make some additional judgments. Be sure to complete this portion as well.

Part I: An important part of this study is the grader's reaction to the instruments he is using to evaluate the papers. Please write as fully as you can your appraisal of the following aspects of the instrument you used in Trial 2, after answering each question with a "Yes" or "No".

1. Adequacy for measuring competence in competence:

a. Does it meet your standards for a logical, coherent, reliable tool for evaluating composition?

Yes_____
No _____

b. Does it provide for the qualities you expect in good writing?

Yes_____
No _____

c. Is there anything omitted that should be included?

Yes_____
No _____

d. Is it properly weighted in its parts?

Yes_____
No _____

2. Clarity of the directions for using the instrument:

Is there any portion of this instrument that is unclear or ambiguous?

Yes_____
No _____

3. Efficiency and ease in evaluating a set of papers:

a. Was the instrument easy to use?

Yes_____
No _____

(Continue to Page 2)

b. Do you think it was easier to grade the paper with this instrument than with the instrument you used in Trial 1?

Yes_____
No _____

c. Do you think it was easier to grade the paper with this instrument than with no instrument?

Yes_____
No _____

d. Do you think using the instrument decreased the time you would normally take to grade a paper?

Yes_____
No _____

e. Do you think using the instrument increased the time you would normally take to grade a paper?

Yes_____
No _____

Part II: Please add your answers to the following questions to your answers for Part I. Use the backs of these sheets or attach additional sheets with your grader number on them.

1. Suggestions for revisions:

What are your suggestions for revising:

a. The first scale?
b. The second scale?
c. This study?

2. Grader's choice:

Which of the following choices would you prefer for evaluating a set of compositions:

a. The first scale?
b. The second scale?
c. No instrument?
d. An instrument other than(a) or (b)?

State briefly the reasons for your answer.

3. Comparison of Scale 1 and Scale 2:

Now that you have used both instruments, what do you think are the strengths and weaknesses of each instrument?

VIII. Report on Comparison of Two Instruments Pilot Study, December, 1976

TO: Members of the English Department

FROM: Alice Barter

DATE: 2 December 1976

RE: Report on Comparison of Two Instruments for the Evaluation of Composition

Attached is a report on the results of an initial study in the comparison of evaluative instruments for composition. I wish to express here my gratitude to Donna Skuta, Richard Rose, Virginia Gibson, Mary Rose Shaughnessy, Dan Stone, Charles Totten, James Doppke and Dr. Victor Pellegrini of the Computer Services for their contributions in moving the work along at its various stages.

I hope the report will be of some interest to you. It follows this sequence:

I Introduction

II Findings;
- Agreement
- Time
- Frequencies-Error Count
- Correlations
- Grader's Comments

III Conclusion

REPORT ON THE COMPARISON OF TWO INSTRUMENTS

FOR THE EVALUATION OF COMPOSITION

I. INTRODUCTION:

Two experimental forms of an instrument for evaluating composition were developed by me in 1975, Form A, to be used in a regular testing program for competence in composition, and Form B, to be used in judging contests in composition. (A copy of Form A is attached to this report.) Form B differs from Form A in having an additional category, Originality, in Part I. It has not been used to date.

In April, 1976, four full-time department members who volunteered and one part-time faculty member selected by Rosemary Hake participated in the first phase of this study (Trial 1). They were Virginia Gibson, Mary Rose Shaughnessy, Charles Totten, Daniel Stone, and Richard Rose. Each except for one who graded only 14 papers was asked to grade 20 papers using the departmental instrument (Hake). In addition, he was to record the time he took to grade each paper, to write a brief appraisal of the adequacy, clarity and efficiency of the instrument and to offer any suggestions he had for revising it. Six weeks later, using Form A, he was asked to re-evaluate xeroxed copies of the same papers (Trial 2) and follow the same instructions for time-keeping, appraisal and comments as he had used in Trial 1. Copies of these comments by each grader for Trials 1 and 2 are attached to some reports. Because some of the graders did not follow directions or misunderstood them, some data were missing in both trials. These errors and omissions were accurately recorded and accounted for in working with the data.

In spite of these restrictions and the fact that the xeroxed copies sometimes showed previous corrections, relevant and, I believe, useful informations was derived from this experiment. Guidelines for the study were the following questions:

(1) Is there a difference between the Pass or Fail judgment on Trial 1 and Trial 2? If so, to what degree?

(2) Using the experimental instrument for Trial 2, is there a difference in the time for grading as reported for Trial 1 and Trial 2? If so, how much?

(3) What are the graders' reactions to the clarity, adequacy, and efficiency of each instrument?

The data were also analyzed for another question, "As indicated by the scores, what relationships, if any, are there between

(a) the parts of each instrument to the Pass or Fail grade,

(b) the parts of an instrument to each other;

(c) the parts of the Hake instrument to parts and whole of the experimental one?"

The question will be dealt with more carefully in Experiment 2 the findings are given here for this experiment since they are available.

The experiment will be repeated as soon as it is feasible with the instruments in their present form but with some variation in procedure:

(1) the papers to be read will be those of incoming freshmen or transfer students who have not been trained to be tested by the present instrument

(2) the graders will be graduate students in English, or other qualified persons, who have not previously used either instrument.

(3) instead Of 5 graders at least 10, and if possible 20, will be used, The results of the two experiments will be compared and reported to the department.

II. FINDINGS:

There were 15 variables in this study:

(1) student number
(2) grader number
(3) Part I-A
(4) Part I-B
(5) Part I-C
(6) Part II
(7) Pass/Fail-Trial 2
(8) Total Score-Trial 2
(9) Time-Trial 2
(10) Pass/Fail Trial 1
(11) Dimension 1
(12) Dimension 2
(13) Dimension 3
(14) Dimension 4;
(15) Time-Trial 1

The data were analyzed for frequency of occurrence, means and degrees of correlation of some variables with others. The latter were of secondary interest to this study. Analysis was done in three ways, for the general population, for the sub-populations of Pass and Fail, and by grader. In the information that follows Trial 1 designates the use of the Hake instrument and Trial 2 the use of the Barter instrument. The total number of students

involved was 94.

A. Agreement

1. The frequency of Pass or Fail grades on each Trial were as follows:

   (a) In Trial 1 fifty-one students [54.3%] failed while 43 [45.7%] passed. There were no missing observations.

   (b) In Trial 2, 71 students [75.5%] failed while 23 [24.5%] passed. There were no missing observations.

2. Agreement of Pass or Fail between Trial 1 and Trial 2 occurred as as follows:

   (a) Agreement occurred in 68 casses or 72.3% of the time.

   (b) There was disagreement in 26 cases or 27.7% of the time. In these 26 cases, 23 grades (88.4%) were changed from Pass on Trial 1 to Fail on Trial 2. Three grades (11.6%) were changed from Fail on Trial 1 to Pass on Trial 2.

   (c) The average number of grade changes for the five graders was 5.20. By grader they occurred as follows:

Grader	Number of Changes
1	5 out of 20 cases
2	6 out of 20 cases
3	5 out of 20 cases
4	5 out of 14 cases
5	5 out of 20 cases

B. Time

In both Trial 1 and Trial 2 more than 60 per cent of the papers were graded in 5 minutes or less (Trial 1-63.9%) (Trial 2-64.9%). The time it took to grade the papers in Trial 1 and Trial 2 is shown below:

1. Trial 1

Only 36 (28.3%) of these papers designated the time for grading. The mean time for this group was 5.35 minutes. Grader 1 gave no information for the 20 papers graded. Grader 2 gave the time for only 4 papers out of 20: of the 3 Fail papers, the mean time for grading was 8.33 minutes; the time for the one Pass paper was 9.00 minutes. The mean time for Grader 2 on the 4 papers was 8.50. Grader 3 gave information on 18 out of 20

papers. For the 11 Fail papers the mean time was 4.73; for the 7 Pass papers it was 3.64 minutes. The mean time for Grader 3 on the 18 papers was 4.31 minutes. Grader 4 gave information for all 14 papers he graded. For the 7 Fail papers the mean time was 6.00 minutes; for the 7 Pass papers it was 5.57 minutes. Grader 5 provided no information for the 20 papers he graded.

2. Trial 2

Information about time was provided for all 94 papers in Trial 2. The mean time for the total population on Trial 2 was 5.17 or slightly less than for Trial 1. Grader 1: For the 11 Fail papers out of 20 the mean time was 6.81 minutes; for the 9 Pass papers it was 5.11 minutes. The mean time for this grader was 6.05 minutes on 20 papers. Grader 2: For the 19 out of the 20 Fail papers the mean time was 6.53 minutes; for the 1 Pass paper it was 10.00 minutes. The mean time for this grader on 20 papers was 6.70 minutes. Grader 3: For the 15 Fail out of the 20 papers, the mean time was 3.67 minutes; for the 5 Pass papers the mean time was 3.40 minutes. The mean time for this grader on 20 papers was 3.60 minutes. Grader 4: For the 10 Fail papers out of 14 the mean time was 5.60 minutes; for the 4 Pass papers it was 4.50 minutes. The mean time for this grader on 14 papers was 5.29 minutes. Grader 5: For the 16 Fail papers out of 20 the mean time was 4.31 minutes; for the 4 Pass papers it was 4.00. The mean time for this grader on 20 papers was 4.25 minutes.

C. Error Count

1. Trial 1

(a) On the Trial 1 instrument the mean error count for the total population was as follows:

D1 = 1.57 D2 = 4.05 D3 = 8.06 D4 = 3.01

(b) The mean error count on Dimensions 1,2,3,4 by grader and Pass or Fail was as follows:

	Mean-D-1	Mean-Pass	Mean-Fail
Grader 1	.550	.143 (N=14)	1.50 (N=6)
Grader 2	1.00	.143 (N=7)	1.46 (N=13)
Grader 3	1.60	.750 (N=8)	2.17 (N=8)
Grader 4	1.14	.286 (N=7)	2.00 (N=7)
Grader 5	3.45	.857 (N=7)	4.87 (N=13)

	Mean-D-2	Mean Pass	Mean Fail
Grader 1	3.15	2.07 (N=14)	5.67 (N=6)
Grader 2	4.20	2.43 (N=7)	5.15 (N=13)
Grader 3	4.35	1.87 (N=8)	6.00 N=12)
Grader 4	5.14	*4.14 (N=7)	6.14 (N=7)
Grader 5	3.75	2.29 (N=7)	*4.54 (N=13)

	Mean-D-3	Mean Pass	Mean-Fail
Grader 1	7.65	5.93 (N=14)	11.67 (N=6)
Grader 2	12.20	*8.00 (N=7)	14.46 (N=13)
Grader 3	6.00	5.00 (N=8)	* 6.67 (N=12)
Grader 4	7.57	6.00 (N=7)	9.14 (N=7)
Grader 5	6.75	1.29 (N=7)	9.69 (N=13)

	Mean-D-4	Mean Pass	Mean-Fail
Grader 1	4.15	4.57 (N=14)	*3.17 (N=6)
Grader 2	4.35	3.57 (N=7)	4.77 (N=13)
Grader 3	3.70	3.00 (N=8)	4.17 (N=12)
Grader 4	2.00	*2.00 (N=7)	*2.00 (N=7)
Grader 5	.55	* .71 (N=7)	* .46 (N=13)

BEST COPY AVAILABL

(c) On the scoring of Dimension 1, zero occurred 41 (43.6%) out of 94 cases, 1 occurred in 24 cases (25.5%) 2, occurred in 10 cases (10.6%), the range of the error count being from a low of zero to a high of 16. The median error count for Dimension 1 fell between 0 and 1. There were no missing cases.

(d) On the scoring of Dimension 2, the range was from a low of zero to a high of 13. The median error count fell between 3 and 4. There were no missing cases.

(e) On the scoring of Dimension 3 the range of the error count was from zero to 32. The median fell between 6 and 7. There were no missing cases.

(f) On the scoring of Dimension 4 the range of the error count was from zero to 12. The median error count fell between 1 and 2, one being at 33% and two being at 53.0%. There were no missing cases.

2. Trial 2

In the scoring of Part I-Trial 2, a score of zero would mean that the grader found no flaws in Part I. A score of 8 would be the maximum penalty for Part I (A=4, B=2, C=2). The following occurred:

0	occurred 12 times	(12.8%)
2	occurred 13 times	(13.8%)
4	occurred 11 times	(11.7%)
6	occurred 5 times	(5.3%)
8	occurred 22 times	(23.4%)

Those papers that were weighted incorrectly (i.e.,not by the scale) were recorded as missing observations for that portion of the test. They were also not counted in the results for Total Score-Trial 2.

Because the mean includes zero scores in Part I, A, B, C, it tends to be much lower than the weighting of a part score.

(a) On the Trial 2 instrument the mean error count for the total population on the parts of the instrument was as follows:

Part I-A = 1.68 (N=69)
Part I-B = 1.30 (N=69)
Part I-C = 1.22 (N=74)
Part II = 10.65 (N=94)
Total Score Trial 2=20.36 (N=63)

(b) The mean error count on Part I, A, B, C, Part II and Total Score by grader and Pass or Fail was as follows:

BEST COPY AVAILABLE

	Mean-Part I-A	Mean-Pass	Mean-Fail
Grader 1	Data incomplete		
Grader 2	2.60	.00 (N=1)	.80 (N=15)
Grader 3	2.60	.80 (N=5)	3.20 (N=15)
Grader 4	.67	.00 (N=4)	1.00 (N= 8)
Grader 5	2.00	.00 (N=4)	2.50 (N=16)

	Mean-Part I-B	Mean-Pass	Mean-Fail
Grader 1	Data incomplete		
Grader 2	.95	.00 (N=1)	*1.11 (n=18)
Grader 3	1.20	.00 (N=5)	1.60 (N=15)
Grader 4	1.56	*1.00 (N=2)	1.71 (N= 7)
Grader 5	1.60	.00 (N=4)	2.00 (N=16)

	Mean-Part I-C	Mean-Pass	Mean-Fail
Grader 1	Data incomplete		
Grader 2	1.00	.00 (N=1)	*1.05 (N=19)
Grader 3	1.40	*1.20 (N=5)	1.47 (N=15)
Grader 4	.77	* .50 (N=9)	* .89 (N= 4)
Grader 5	1.50	.00 (N=4)	1.87 (N=16)

	Mean-Part II	Mean-Pass	Mean-Fail
Grader 1	17.50	12.89 (N=9)	21.27 (N=11)
Grader 2	25.05	*15.00 (N=1)	25.58 (N=19)
Grader 3	12.95	6.80 (N=5)	*15.00 (N=15)
Grader 4	12.07	6.50 (N=4)	14.30 (N=10)
Grader 5	14.30	9.00 (N=4)	15.63 (N=16)

	Mean-Total Score	Mean-Pass	Mean-Fail
Grader 1	Data incomplete		
Grader 2	26.73	15.00 (N=1)	27.51 (N=14)
Grader 3	18.15	8.80 (N=5)	21.27 (N=15)
Grader 4	15.00	9.00 (N=2)	17.40 (N= 5)
Grader 5	19.40	9.00 (N=4)	22.00 (N=16)

*Since 75.5% failed on this instrument, the mean total error count for the total population (without separation into Pass and Fail is closer to the means of the Fail group.

(c) On the scoring of Part I-A zero (no error) occurred in 40 (42.6%) out of 69 cases, while 4 occurred in 29 cases (30.9%), and there were 25 missing cases (26.6%).

(d) On the scoring of Part I-B zero occurred in 24 (25.5%) out of 69 cases, while 2 occurred in 45 (47.9%) cases, and there were 25 (26.6%) missing cases.

(e) On the scoring of Part I-C zero occurred in 29 (30.9%) out of 74 cases, while 2 occurred in 45 (47.9%) cases, and there were 20 (21.3%) missing cases.

(f) The number of errors found in Part II ranged from the low of 2 to the high of 50. The median error count in Part II was 16.5.

(g) The total error count on Trial 2 ranged from a low of 2 to a high of 44.*

*This number is lower than that for the sub-total, Part II, because some papers with errors in the scoring of Part I are listed as the 31 missing cases cases. Only the 63 valid cases were used.

The median error count for the total score on Trial 2 was 19.5.

D. Correlations

1. The correlations of the Pass/Fail variable with scores for D-1. D-2, D-3 were very significant (all were at the .001 level of significant), but there seemed to be no relationship between D-1 and the Pass/Fail variable (at .840 level of significance).

2. The correlations of the Pass/Fail variable with scores on all parts of the Trial 2 instrument were very significant: Part I-A (.003): Part I-B (.001) Part I-C (.006) Part I (A,B,C) combined) (.016) Part II (.001) Total Score (.001).

3. The correlations of scores on Part I-Trial 2 with those on Dimensions I and 2 for the total population was as follows:
   a. There was a significant relationship between scores on Dimension 1 and Part I-A (.001) Part I-B (.006), Part IC (.001) and those on Part I (A,B,C combined) (.012)

   b. The relationships between Dimension 2 and Part I(A, B and C combined) or the Total Score-Trial 2 were notsignificant.

c. The relationship between Dimension 3 and the Total Score-Trial 2 was very significant (.001) as was that between Dimension 4 and the Total Score-Trial 2 (.02).

4. The correlations of the scores on the parts of the Trial 1 instrument with each other were significant as follows:

a. For the Fail group-Trial 1

D-1 to D-3= .02 D-1 to D-4= .03
There was no significant relationship between D-2 and either D-1, D-3, or D-4.

b. For the Pass group-Trial 1

D-3 to D-4= .001

There were no significant relationships between the other dimensions for this group.

c. For the Fail group-Trial 2

D-1 to D-4= .02 D-2 to D-3 = .03

There were no significant relationships between the other dimensions for this group.

d. For the Pass group- Trial 2

D-2 to D-3= .05
There were no significant relationships between the other dimensions for this group.

5. The correlations of the scores of parts of the Trial 2 instrument with each other were significant as follows:

a. For the Fail group-Trial 1

Part I-A to Part I-C= .009 Part I-C to Total - Part I= .04
Part I-B to Part I-C= .03 Part II to Total Score= .001

b. For the Pass group- Trial 1

Part I-A to Part I-C= .004 Part I-Total to Part II= .05
Part I-Total to Total Score= .001

c. For the Fail group-Trial 2

Part I-A to Part I-C= .001 Part I-C to PartI-Total=.01
Part I-B to Part I-C= .03 Part II to Total Score= .001
Part I-B to Part II= .02

d. For the Pass group Trial 2

Part I-Total to Part II = .03
Part II to Total Score = .001

6. The correlations of the scores on Part I-A, B, C, and Total-Part I with those of Dimensions 1 and 2 were significant as follows:

a. For the Fail group-Trial 1
Part I-A to D-1 = .009
Part I-B to D-1 = .05
Part I-C to D-1 = .03

There was no significant relationship to D-2 nor between Total-Part I and D-1.

b. For the Pass group-Trial 1
Part I-Total to D-1 = .02

All other relationships were not significant for this group.

c. For the Fail group-Trial 2
Part I-A to D-1 = .001
Part I-C to D-1 = .005

All other relationships were not significant for this group.

d. For the Pass group=Trial 2

There were no significant relationships in this group.

7. The correlations of the scores on Part II with those of D-3 and D-4 were significant as foldows:

a. For the Fail group-Trial 1
Part II to D-3 = .001
Part II to D-4 = .04

b. For the Pass group-Trial 1
Part II to D-3 = .001
Part II to D-4 = .001

c. For the Fail group-Trial 2
Part II to D-3 = .001
Part II to D-4 = .003

d. For the Pass group-Trial 2
Part II to D-3 = .012
Part II to D-4 = .004

8. The correlation of Total Score-Trial 2 with the scores of D-3 and D-4 were significant as follows:

   a. For the general population
   Total Score to D-3 = .001
   Total Score to D-4 = .02

   b. For the Fail group-Trial 1
   Total Score to D-3 = .001
   The relationship to D-4 was not significant

   c. For the Pass group-Trial 1
   Total Score to D-3 = .001
   Total Score to D-4 = .007

   d. For the Fail group-Trial 2
   Total Score to D-3 = .001
   The relationship to D-4 was no significant

   e. For the Pass group-Trial 2
   There was no significant relationship for this group.

9. The correlations of Time for correction between Time 1 and Time 2 were significant as follows:

   a. For the general population = .001

   b. For the Fail group-Trial 1 = .003

   c. For the Pass group-Trial 1 = .001

   d. For the Fail group-Trial 2 = .001

   e. For the Pass group-Trial 2 = .005

E. Summary of Graders' Comments

Copies of each grader's remarks in full about each instrument are attached to some of the reports. Briefly, the graders generally seemed satisfied with the Hake instrument, citing as its particular virtues its reliability, specificity and its value as a diagnostic tool. One grader noted that while he was a fervent admirer of the instrument he still did not understand how to use it and suggested some revisions in the wording of it. Criticism of it dealt with inadequacy (for illogicality, incoherence), difficulty of use, lack of clarity and inefficiency. Comments

about the Barter instrument tended to be more diverse. Its advantages were that it was easier to use, it allowed for more than one type of organizational pattern, it was efficient, and the practice of circling the errors helped to emphasize them. Its disadvantages were that what was saved in "mechanical time" was used in "judgmental time", it was not a good diagnostic/teaching tool, it lacked specificity, and there was disagreement about the weighting of errors in Part I.

III. Conclusion

In answer to the questions posed at the beginning of the study, the results show that there was a sizable difference between the grading judgments of Trial 1 and Trial 2 and that the change in judgment usually was from Pass to Fail. The average number of changes per grader was five or a little more than 25%. Why these changes occurred, whether they were a function of the grader's attitude toward an unfamiliar instrument or that the second instrument gave a clearer delineation of the grader's standards, or that there are other reasons, is not clear from a grader's own comments; the question will need further study. The results also show - , contrary to the researcher's expectations that the instrument used for Trial 2 would save time - the time for grading was nearly the same in the two trials. Two possible explanations of this result are 1) that the time saved in not denoting the flaws by dimensions was used in weighting the flaws in Part I, and 2) that the data for this question are unreliable because not all graders followed directions in keeping records of the time. It should be noted that half of the data for Time-Trial 1 was provided by Grader 3 who had the lowest average for Time - Trial 2. In the second experiment directions to the graders will need to be given both orally and in written form; there should also be a direct question about why he thinks time was saved or not saved in the second trial. (The order of using the instruments will be reversed for Experiment 2).

Barter

The graders' comments showed some confusion about the purpose of the experiment and the attitude that neither instrument satisfied the grader's concept of an adequate measuring device. There was a tendency to feel more confident about the Hake instrument, citing as its particular merits its specificity and its use both as a diagnostic and a teaching tool. The evaluations of the Trial 1 instrument were written before the graders had used the one for Trial 2. In one case, because his original had been misplaced, the grader was asked to write his comments for Trial 1 after he had used both instruments. The original was located later and is included; it is interesting to compare the two versions, particularly their length and the additional comment about how the conditions under which the experiment was conducted tended to create a bias in favor of the Hake instrument. Another grader expressed at length his frustrations over the low quality of what the Trial 1 instrument is designed to measure, that is, that it doesn't get at the stupidities, misinformation, or meagerness of information that mar student writing.

Comments on the Trial 2 instrument are harder to classify. In general, graders found it more flexible and easier to use but not adequate for diagnosis or teaching. Reactions toward Part I-A,B,C were that they allowed for qualities that were omitted in the Trial 1 instrument, that there was some overlap between Part I-A and Part I-B, that errors in Part I should be

given particular attention in the English course, that a grader could predict whether a paper was Pass or Fail by the time he had graded Part I. Questions were raised as to whether the quality of complexity was a virtue in writing, the disproportionately small number of errors counted for flaws in organization, logic and complexity, and how it would be possible to set standards for grading using this weighting system. One grader suggested that the idea of circling the errors in Part II be adopted whatever instrument was used by the department. Also, it was pointed out by one grader that if the purpose of the grading instrument is merely to arrive at a Pass/Fail judgment, the Trial 2 instrument seemed to have a slight edge, but if it is to be used instructionally, then the Trial 1 instrument, with some revisions, is superior because of its specificity.

As stated earlier, the correlations tests that were run provided data that go beyond the purpose of Experiment 1. Whether there is a relationship between what the parts and whole of the two instruments are measuring and whether these relationships can be validly measured by scores will be tested more fully in Experiment 2.

IX. Notes

NOTES

1. National Council of Teachers of English, "Common Sense and the Teaching of English" (Urbana:1975), p.7.

2. Richard Braddock, "Evaluation of Writing Tests," in Reviews of Selected Published Tests in English,ed. Alfred H. Grommon (Urbana: National Council of Teachers of English, 1976), p. 118.

3. John C. Mellon, National Assessment and the Teaching of English (Urbana: National Council of Teachers of English, 1975). See also Alfred H. Grommon, Reviews of Selected Published Tests in English (Urbana: NCTE, 1976).

4. Paul Diederich, Measuring Growth in English(Urbana:NCTE, 1974). See also Rosemary Hake and David Andrich,"The Ubiquitous Essay: A Discourse and Psychometric Model to Identify, Measure, Evaluate, and Teach Writing Ability"(Unpuclished Monograph, University of Western Australia, 1974).

5.Rosemary Hake and Jesse Green, "A Test to Teach to: Composition at Chicago State University," ADE Bulletin, 52,(February, 1977), pp.3-7. See also Rosemary Hake and David Andrich, "Time to Write Project," ESEA,IVC, Report of Third Evaluation,"(1980), p.5.

6. Lee Odell and Charles R. Cooper, "Procedures for Evaluating: Writing Assumptions and Needed Research," College English, 42 (September, 1980). See also: Anne R. Gere,"Written Compositions: Toward a Theory of Evaluation," in the same issue.

7. Alice Barter, "A Report on the Comparison of Two Instruments for the Evaluation of Composition." Unpublished report to the Department of English and Speech, Chicago State University, December, 1976. A copy of this report may obtained by writing to the author.

REFERENCES

References

1. Andrich, David and Rosemary Hake. "The Application of Discourse Theory and a Rasch Model for Measuring in the Evaluation of Written Expression". Educational Research and Perspectives, 1 (December, 1974), 51-61.

2. Barter, Alice. "A Report on the Comparison of Two Instruments for the Evaluation of Composition." Unpublished report to Department of English, Chicago State University, December, 1976.

3. Braddock, Richard. "Evaluation of Writing Tests". In Reviews of Selected Tests in English. Ed. Alfred H. Grommon. Urbana, National Council of Teachers of English, 1976.

4. Breland, Hunter. "Multiple Choice Test Assesses Writing Ability," Findings, 4, No. 1 (Fall, 1977), 1-4. Educational Testing Services.

5. Britton, James et al., The Development of Writing Abilities, 11-18. London: Schools Council Research Studies, 1975.

6. Center for Academic Skills. "Alternative Approaches to the Teaching of Writing Skills." City College of New York, 1978.

7. Cooper, Charles R. and Lee Odell. Evaluating Writing. Urbana: National Council of Teachers of English, 1977.

8. Diederich, Paul B. Measuring Growth in English. Urbana: National Council of Teachers of English, 1974.

9. Fagan, William T., Charles R. Cooper and Julie M. Jensen. Evaluating in the English Language Arts. Urbana: National Council of Teachers of English, 1975.

10. Gere, Anne Ruggles. "Written Composition: Towards a Theory of Evaluation." College English, 42 (September, 1980), 41-48, 53-58.

11. Godshalk, Fred I., Frances Swineford, and William E. Coffman. The Measurment of Writing Ability. New York: College Entrance Examination Board, 1966.

12. Hake, Rosemary and David Andrich. "The Ubiquitous Essay: A Discourse and Psychometric Theory Essay Writing Ability". Australia, 1974.

13. Hake, Rosemary and David Andrich. "Time to Write Project," ESEA, IV C, Report of Third Year Evaluation", 1980.

14. Hake, Rosemary and Jesse D. Green. "A Test to Teach To: Composition at Chicago State University." ADE Bulletin, 52 (February, 1977), 31-37.

15. Larson, Richard. "The Whole Is More Than the Sum of Its Parts". Arizona English Bulletin, 16 (February, 1974), 175-181.

16. Mellon, John C. National Assessment and the Teaching of English. Urbana: National Council of Teachers of English, 1975.

17. National Council of Teachers of English. "Common Sense and Testing in English." Urbana, 1975.

18. O'Dell, Lee and Charles R. Cooper. "Procedures for Evaluating Writing: Assumptions and Needed Research." College English, 42 (September, 1980), 35-43.

19. Steel, R.G.D. and J.H. Torrie. Principles and Procedures of Statistics. New York: McGraw Hill, 1960.

BEST COPY